マルチスケールディープネットワークを用いた単一画像からの深度マップ予測

David Eigen[1]
deigen@cs.nyu.edu

Christian Puhrsch[1]
cpuhrsch@nyu.edu

Rob Fergus[1,2]
fergus@cs.nyu.edu

[1]Dept. of Computer Science, Courant Institute, New York University
[2]Facebook AI Research

Abstract

奥行きの予測は、シーンの3Dジオメトリを理解する上で不可欠な要素である。ステレオ画像では局所的な対応関係で推定に十分であるが、1枚の画像から奥行き関係を求めるのは容易ではなく、様々な手がかりから大域的な情報と局所的な情報の両方を統合する必要がある。さらに、このタスクは本質的に曖昧であり、全体的なスケールから来る大きな不確実性の原因である。本論文では、2つのディープネットワークスタックを採用することで、このタスクに対処する新しい方法を提示する。1つは、画像全体に基づいて粗いグローバル予測を行うものであり、もう1つは、この予測を局所的に洗練するものである。また、スケールではなく、スケール不変誤差を適用して、奥行き関係の測定に役立てる。生データセットを大規模な学習データ源として活用することで、本手法はNYU DepthとKITTIの両方で最先端の結果を達成し、スーパーピクセル化を必要とせずに詳細な深度境界をマッチングする。

1 Introduction

奥行きの推定は、シーン内の幾何学的関係を理解する上で重要な要素である。その結果、このような関係は、物体やその環境のより豊かな表現を提供するのに役立ち、しばしば既存の認識タスク[18]の改善につながるだけでなく、3Dモデリング[16, 6]、物理・支援モデル[18]、ロボット工学[4, 14]、潜在的にオクルージョンに関する推論など、多くのさらなる応用を可能にする。

ステレオ画像や動きに基づく奥行き推定に関する先行研究は多いが[17]、1枚の画像から奥行きを推定する研究は比較的少ない。しかし、実際には単眼的なケースがしばしば生じる: 潜在的な応用例としては、ウェブやソーシャルメディアのアウトレット、不動産のリスティング、ショッピングサイトに配信される多くの画像の理解を深めることなどが挙げられる。これらは屋内と屋外の両方のシーンの例を多く含んでいる。

単眼の場合、ステレオの場合と同程度にまだ取り組まれていない理由はいくつか考えられる。正確な画像対応があれば、ステレオの場合、奥行きを決定論的に復元することができます[5]。このように、ステレオ奥行き推定は、ロバストな画像点対応関係の開発に還元することができ、局所的な外観特徴を用いて見つけることができることが多い。対照的に、1つの画像から奥行きを推定するには、線の角度や遠近感、物体の大きさ、画像の位置、大気の影響など、単眼の奥行き手がかりを使用する必要がある。さらに、これらを効果的に関連付けるためには、シーンの全体像が必要かもしれないが、ステレオには局所的な視差で十分である。

さらに、このタスクは本質的に曖昧であり、技術的に非論理的な問題である:ある画像が与えられたとき、無限の可能な世界シーンがそれを作り出したかもしれない。もちろん、これらのほとんどは実世界の空間では物理的にありえないものであり、したがって、深さはかなりの精度で予測される可能性がある。しかし、少なくとも一つの大きな曖昧さが残っている。極端なケース(通常の部屋か人形屋か)はデータに存在しないが、部屋や家具の大きさには中程度のばらつきがある。

より一般的なスケール依存の誤差に加えて、スケール不変の誤差を用いてこれに対処する。これは、一般的なスケールではなく、シーン内の空間的な関係に注目するものであり、特に3Dモデリングのような、後処理中にモデルが再スケールされることが多いアプリケーションに適している。

本論文では、1枚の画像から奥行きを推定するための新しいアプローチを紹介する。まずシーンの大域的な構造を推定し、次に局所的な情報を用いてシーンを洗練させる、という2つの要素を持つニューラルネットワークを用いて、奥行きに直接回帰する。ネットワークは、ポイントワイズエラーに加えて、ピクセル位置間の深度関係を明示的に考慮した損失を用いて学習される。本システムは、NYU DepthとKITTIにおいて、最先端の推定率を達成し、定性的な出力も改善された。

2 Related Work

我々の研究に直接関連するのは、1つの画像から奥行きを推定するいくつかのアプローチである。Saxenaら[15]は、線形回帰とMRFを用いて画像特徴のセットから奥行きを予測し、後に3Dモデル生成のためのMake3D[16]システムに彼らの研究を拡張している。しかし、このシステムは画像の水平アライメントに依存しており、制御の少ない設定に悩まされている。Hoiemら[6]は、奥行きを明示的に予測するのではなく、画像領域を幾何学的構造(地面、空、垂直)に分類し、これを用いてシーンの単純な3Dモデルを構成している。

より最近では、Ladickyら[12]が、性能を向上させるために、意味的なオブジェクトラベルを単眼的な奥行き特徴と統合する方法を示している。しかし、彼らは手作りの特徴に依存し、画像をセグメント化するためにスーパーピクセルを使用している。Karschら[7]は、SIFT Flow[11]に基づくkNN転送メカニズムを用いて、単一画像から静的背景の深さを推定し、それを動き情報で補強することで、動画中の動く前景被写体をより良く推定する。これにより、より良いアライメントを実現することができるが、データセット全体が実行時に利用可能であることが必要であり、高価なアライメント手順を実行する。これに対して、本手法は、より記憶しやすいネットワークパラメータのセットを学習し、リアルタイムで画像に適用することができる。

より広範に、ステレオ奥行き推定が広く研究されている。Scharsteinら[17]は、マッチング、集約、最適化技術によって整理された、2フレームステレオ対応のための多くの方法の調査と評価を提供している。マルチビューステレオの創造的な応用として、Snavelyら[20]は、同じシーンの多くの未校正の消費者写真のビューを横断的に照合し、共通のランドマークの正確な3D再構成を作成する。

機械学習技術はステレオケースにも適用されており、注意深いカメラアライメントの必要性を緩和しながら、しばしばより良い結果を得ている[8, 13, 21, 19]。この研究に最も関連するのはKondaら[8]で、ステレオシーケンスから奥行きを予測するために画像パッチ上で因数分解オートエンコーダを訓練する。

また、単一画像の奥行き推定には、ハードウェアを用いたソリューションがいくつかある。Levinら[10]は、修正されたカメラアパーチャーを使用してデフォーカスから深度を実行し、KinectとKinect v2は、アクティブステレオと飛行時間を使用して深度をキャプチャする。本手法では、このようなセンサを間接的に利用することで、学習時にグランドトゥルースの奥行き目標を提供する。しかし、テスト時には、本システムは純粋にソフトウェアベースで、RGB画像から奥行きを予測する。

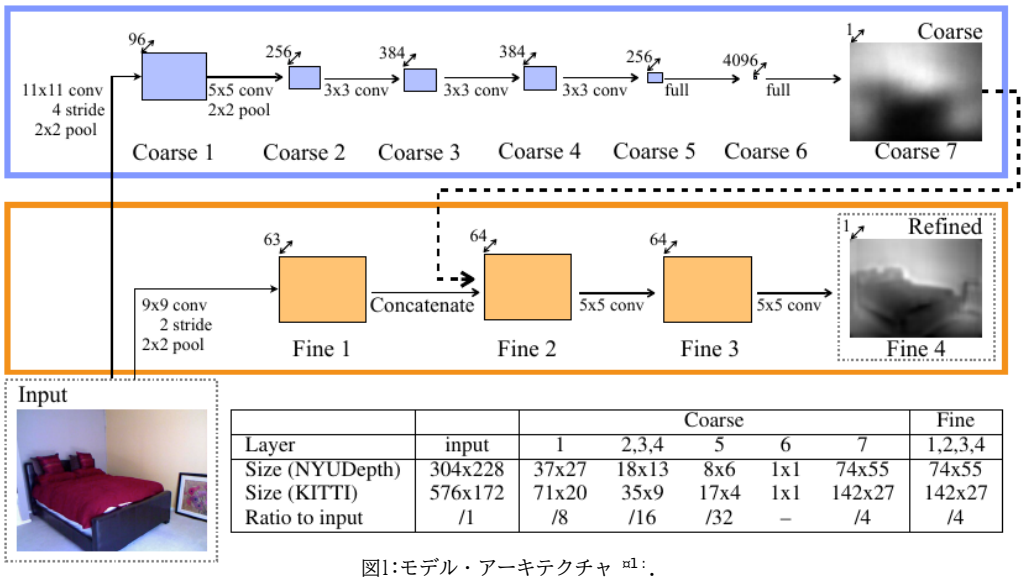
3 Approach

3.1 モデルアーキテクチャ

我々のネットワークは、図1に示すように、2つのコンポーネントスタックで構成されている。粗いスケールのネットワークは、まず、シーンの深さをグローバルなレベルで予測する。そして、これを局所領域内でファインスケールネットワークにより精緻化する。両スタックは元の入力に適用されるが、さらに粗いネットワークの出力は、追加の第1層画像特徴として細かいネットワークに渡される。このようにして、ローカルネットワークは、より細かいスケールの詳細を取り込むために、グローバル予測を編集することができる。

3.1.1 大域的粗視化ネットワーク

粗いスケールのネットワークのタスクは、シーンのグローバルビューを使用して、全体的な深度マップ構造を予測することである。このネットワークの上位層は完全に接続されているため、視野内に画像全体が含まれる。同様に、下層と中層は、最小プール演算によって画像の異なる部分からの情報を小さな空間次元に結合するように設計されている。そうすることで、ネットワークはシーン全体のグローバルな理解を統合し、奥行きを予測することができる。



		Coarse					Fine
Layer	input	1	2,3,4	5	6	7	1,2,3,4
Size (NYUDepth)	304x228	37x27	18x13	8x6	1x1	74x55	74x55
Size (KITTI)	576x172	71x20	35x9	17x4	1x1	142x27	142x27
Ratio to input	/1	/8	/16	/32	–	/4	/4

図1:モデル・アーキテクチャ [1].

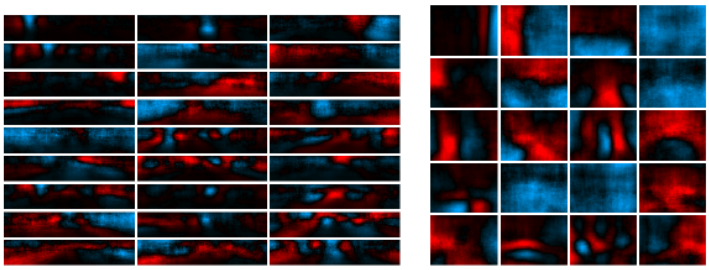
このような理解は、消失点、物体の位置、部屋の位置合わせなどの手がかりを効果的に利用するために、単一画像の場合に必要である。ローカルビュー(ステレオマッチングによく使われる)は、このような重要な特徴に気づくには不十分である。

図1に示すように、グローバルな粗いスケールのネットワークは、畳み込みとマックスプーリングの5つの特徴抽出層と、それに続く2つの完全接続層を含む。入力、特徴マップ、出力サイズは[085]で、図1の深さで予測される。最終的な出力は1/4解像度の画像領域で、入力構造(それ自体[086]ダウンサンプリング)と幾何学的に比較される。 }は、[087]入力のほとんど*More*を含むセンタークロップに対応する(*Ladicky*後述する*etal.*[?]、*show*細かいスケール[088]特徴ネットワークの第1層による*labels*による小さな境界*semantic*領域*object*を失い、画像性能を向上させる。)しかし、それらは手作業で作られた特徴に依存し、使用する
空間[089]/[090]次元_{matedepths}^{segmentthe}の*image.*/*of*出力*static*は_{backgrounds}^{Karsch}より大きい
_{from}^{etal.[? }は、の*akNN*よりも大きい。この[091]ハードコードされた*canachieve*アップサンプリング*better*を*tobetter*出力に制限するのではなく、*moving*特徴量*foreground*マップサイズ*subjects*に制限し、*in*依存する。 }に*this*ハードコードされた*canachieve*アップサンプリング*better*してから、*quaresthe*予測*quiresthe*を*entire*細かい*dataset*ネットワークに渡す、 に対して、*tobe*我々は*available*最上位のフル*and*層が*expensive*テンプレート*alignment*を学習できるようにする(74x55*By* for*contrast,* NYU Depth).*ourmethod*
これらの*learns*は_{anesiaer-to-store}ぼやけた*setof*と予想されるが、*network*は*parameters,*、*and* 8x6予測*nreal – time.*
(上の特徴マップサイズ)のアップサンプリング出力[094]より優れている。サンプル出力の重みを図2に示す。

すべての隠れ層[096]は、*a* 整流と線形の活性化にを使用し、2は粗い出力の例外層7、[097]は線形である。マッチング、ドロップアウト集計、完全連結技術の最適化などに適用される。hidden 創造的なレイヤー6.
アプリケーション Snavely coarse-scaleら[?]のネットワークマッチの畳み込みレイヤー(1-5)098は、多くの未校正のImageNet消費者分類写真タスク[1]で事前学習されたビューである。はランダムに初期化するよりもうまくいったが、[100] 機械差分学習はそれほど大きな技術ではなかった[1] 。はステレオの場合に適用され、しばしば
3.1.2 ローカル102ファインスケール近藤
ネットワークら。
グローバルな透視シーケンスを[103]とた後、これを予測するために、深度ローカルマップ、変位、ステレオによるローカルな精密化を提供することに依存する。このコンポーネントの*also*タスク*several*は、粗予測は、オブジェクト*fromdefocus*や壁*using*エッジのような*depth*局所的な詳細*perform*と整合するように受信する。 fine-scale_{cameraaperature,} network*while* stack*the* は*and*畳み込み[106]層のみからなる*Kinect*。ステレオ}と*and*第1層のプーリングステージ*tocapture*を1つずつ持つ*our*エッジ*method*特徴量
_{makesindirectuse}。
粗いネットワーク*toprovide*は*ground*シーン全体*truth*を見るが、*depth*フィールド*targets*は*during*ビューの*training;*
出力*however,*ユニット*at*の細かい*time*ネットワーク*our*は45x45ピクセルの入力である。畳み込み層は、ターゲット出力サイズで特徴マップに*RGB*適用され、入力スケールの1/4で比較的高解像度の出力を可能にする[2]。

より具体的には、粗い出力は追加の低レベル特徴マップとして入力される。設計上、粗い予測は最初の細かいスケールの層の出力と同じ空間サイズ(プーリング後)であり、

事前学習では、[9]と同様に、2つの隠れ層にドロップアウトを適用し、それぞれ4096 4096 1000出力ユニットを持つ2つの完全連結層を積み重ねる。各トレーニング画像の中央256x256領域からランダムに224x224のクロップを用いてネットワークをトレーニングし、最短辺の長さが256になるようにリスケーリングする。このモデルはILSVRC2012検証セットでトップ5のエラー率18.1%を達成し、画像あたり2つの反転と5つの平行移動で投票する。



(a) (b)

図2:レイヤーCoarse 7(粗い出力)からの重みベクトル、(a)KITTIと(b)NYUDepthの場合。赤はプラス(さらに)、青はマイナス(より近い)、黒はゼロ。重みは一様に選択し、l_2ノルムで降順に示す。KITTIの重みは、道路の両側で深さの変化を示すことが多い。NYUDepthのウェイトは壁の位置や出入り口を示すことが多い。

この2つを連結する(図1のFine 2)。後続の層は、ゼロパディング畳み込みを使用してこのサイズを維持する。

すべての隠れユニットは整流された線形活性化を使用する。最後の畳み込み層は、ターゲットの深さを予測するため、線形である。まず、グランドトゥルースのターゲットに対して粗いネットワークを訓練し、次に粗いスケールの出力を固定したまま細かいスケールのネットワークを訓練する(つまり、細かいネットワークを訓練するとき、粗いネットワークをバックプロパゲートしない)。

3.2 スケール不変誤差

シーンのグローバルスケールは、奥行き予測における基本的な曖昧さである。実際、現在の要素別メトリクスを使用して発生した誤差の多くは、単純に平均深度がどの程度予測されるかによって説明できるかもしれない。例えば、NYUDepthで学習したMake3Dは、対数空間のRMSEを用いて0.41の誤差を得る(表1参照)。しかし、各予測の平均対数深度を対応するグランドトゥルースからの平均で代用するオラクルを使用すると、誤差は0.33に減少し、相対的に20%改善される。同様に、我々のシステムでは、これらのエラー率はそれぞれ0.28と0.22である。したがって、シーンの平均スケールを求めるだけで、全誤差の大部分を占めることになる。

このことに動機づけられ、絶対的なグローバルスケールに関係なく、シーン内の点間の関係を測定するためにスケール不変誤差を使用する。予測深度マップyとグランドトゥルースy^*に対して、それぞれiでインデックスされたn個のピクセルを持つ、スケール不変の平均二乗誤差(対数空間)を次のように定義する。

$$D(y,y^*) = \frac{1}{n}\sum_i (\log y_i - \log y_i^* + \alpha(y,y^*))^2, \quad (1)$$

ここで、$\alpha(y, y^*) = \frac{1}{n}\sum_i (\log y_i^* - \log y_i)$ は与えられた(y, y^*)の誤差を最小化するαの値である。任意の予測値yに対して、e^{α}はグランドトゥルースに最もよく整列するスケールである。yのすべてのスカラー倍数は同じ誤差を持ち、したがってスケール不変である。

この指標を見るための2つの追加的な方法は、以下の等価な形式によって提供される。$d_i = \log y_i - \log y_i^*$ を画素iにおける予測値とグランドトゥルースの差とすると、次式が成り立つ。

$$D(y,y^*) = \frac{1}{n^2}\sum_{i,j}\left((\log y_i - \log y_j) - (\log y_i^* - \log y_j^*)\right)^2 \quad (2)$$

$$= \frac{1}{n}\sum_i d_i^2 - \frac{1}{n^2}\sum_{i,j} d_i d_j = \frac{1}{n}\sum_i d_i^2 - \frac{1}{n^2}\left(\sum_i d_i\right)^2 \quad (3)$$

式2は、出力中の画素i,jのペア間の関係を比較することで誤差を表現する。誤差が小さいためには、予測中の各画素のペアは、グランドトゥルース中の対応するペアと同様の深さの差がなければならない。式3は、この指標を元のl_2誤差に関連付けるが、$\frac{1}{n^2}\sum_{ij} d_i d_j$ という項を追加し、同じ方向であれば間違いをクレジットし、反対であればペナルティを与える。したがって、不完全な予測は、その間違いが互いに一致するとき、より低い誤差を持つことになる。式3の最後の部分は、これを線形時間計算として書き直す。

スケール不変誤差に加えて、セクション4で説明するように、先行研究で提案されているいくつかの誤差メトリクスに従って、我々の手法の性能も測定する。

3.3 学習損失

性能評価に加えて、スケール不変誤差を学習損失として使用することも試した。式3からヒントを得て、サンプルあたりの学習損失を次のように設定する。

$$L(y, y^*) = \frac{1}{n}\sum_i d_i^2 - \frac{\lambda}{n^2}\left(\sum_i d_i\right)^2 \quad (4)$$

ここで、$d_i = \log y_i - \log y_i^*$、$\lambda \in [0, 1]$である。ネットワークの出力はlog yである。つまり、最後の線形層はlogの深さを予測する。$\lambda = 0$ とすると要素ごとの l_2 となり、$\lambda = 1$ とするとスケール不変誤差は正確にはこれらの平均、すなわちλ=0.5を用いると、定性的な出力をわずかに改善しながら、良好な絶対スケール予測が得られることが分かる。

学習中、ターゲット深度マップのほとんどは、特にオブジェクトの境界、窓、鏡面付近で、いくつかの欠損値を持つことになる。我々は、単にマスキングして有効な点のみで損失を評価することで、これらに対処する。すなわち、式4のnを目標深度を持つ画素の数に置き換え、深度値を持たない画素iを除いた和を実行する。

3.4 データの拡張

学習データをランダムなオンライン変換で補強する(NYUDepthの値を示す)[2]：

- 入力とターゲット画像は$s \in [1, 1.5]$でスケーリングされ、深さはsで割られる。
- 回転させる：入力とターゲットを$r \in [-5, 5]$度回転させる。翻訳：入力とターゲットは図1に示すサイズにランダムに切り取られる。色：入力値にランダムなRGB値$c \in [0.8, 1.2]^3$ をグローバルに乗算する。フリップ：入力とターゲットは0.5の確率で水平に反転する。

画像のスケーリングと平行移動は、シーンの世界空間形状を保存しないことに注意してください。これは、スケーリングの場合、奥行き値をスケールsで割ることで簡単に補正できる(画像sをs倍大きくすると、カメラをs倍近く効果的に動かすことができる)。翻訳は簡単には修正できないが(カメラを効果的に変更して奥行き値と相容れないようにする)、表現するシーンがわずかに歪んでいるにもかかわらず、提供した余分なデータがネットワークに利益をもたらすことがわかった。他の変換、フリップ、面内回転は、形状を保存する。テスト時には、スケール1.0の単一センタークロップを使用し、回転や色変換は行わない。

4 Experiments

NYU Depth v2 [18]とKITTI [3]の両方の生バージョンでモデルを学習する。生の分布は、より一般的に使用される小さな分布と同じシーンから収集された多くの追加画像を含み、前処理はない;特に、深度値が存在しない点は塗りつぶされないままである。しかし、我々のモデルはこのようなギャップを処理する自然な能力と、大規模なトレーニングセットに対する要求から、このようなフィッティングなデータソースとなる。

4.1 NYU Depth

NYU Depthデータセット[18]は、Microsoft Kinectカメラを使用してビデオシーケンスとして撮影された464の室内シーンで構成されている。我々は公式のtrain/test分割を使用し、トレーニングに249シーン、テストに215シーンを使用し、これらのシーンの生データを使用してトレーニングセットを構築する。RGB入力は640x480から320x240まで半分にダウンサンプリングされる。深度カメラとRGBカメラは異なる可変フレームレートで動作するため、各深度画像を時間的に最も近いRGB画像と関連付け、1つのRGB画像が複数の深度に関連付けられているフレームを破棄する(このような1対多のマッピングは予測できない)。データセットに提供されたカメラ投影を使用して、RGBと深度のペアを揃える。深度の値がないピクセルは欠落したまま、マスクアウトされる。窓、開いた出入り口、鏡面などによる多くの無効な領域を除去するために、各画像について記録された最小値または最大値に等しい深さもマスクアウトする。

学習セットには120Kのユニークな画像があり、夕方のシーン分布(1シーンあたり1200枚)の後に220Kのリストにシャッフルする。694画像NYU Depth v2テストセット(奥行き値が埋め込まれている)でテストする。サイズ32のバッチでSGDを用いて2Mサンプルの粗いネットワークを学習する。そして、それを固定し、1.5Mサンプルの細かいネットワークを訓練する(すでに訓練された粗いネットワークからの出力が与えられる)。学習率は以下の通りである：粗い畳み込み層1〜5は0.001、粗い完全層6と7は0.1、細かい層1と3は0.001、細かい層2は0.01である。これらの比率は、検証セット(最終評価のためにトレーニングセットにフォールドバック)での試行錯誤によって求められ、すべてのレートのグローバルスケールは5倍に調整された。モメンタムは0.9であった。

4.2 KITTI

KITTIデータセット[3]は、車載カメラと深度センサで走行中に撮影された複数の屋外シーンから構成される。生データの「都市」、「住宅」、「道路」カテゴリーから56シーンを使用する。

[2]For KITTI, $s \in [1, 1.2]$, and rotations are not performed (images are horizontal from the camera mount).

これらはトレーニング用28個とテスト用28個に分けられる。RGB画像はもともと1224x368で、半分にダウンサンプリングしてネットワーク入力を形成する。

このデータセットの深度は、回転LIDARスキャナを使用して異なる時間にキャプチャされた、不規則な間隔の点でサンプリングされている。RGBカメラはスキャナが前方を向いたときに撮影するため、RGB撮影時間に最も近い深度を選択することで、各ピクセルでの衝突を解決する。深度はRGB画像の下部のみで提供されるが、我々は画像全体をモデルに投入し、グローバルな粗いスケールのネットワークに追加のコンテキストを提供する(細かいネットワークはターゲット領域に対応する下部のクロップを見る)。

学習セットは1シーンあたり800枚の画像を持つ。重複を避けるため、車が静止している(加速度が閾値以下)ショットは除外する。左右両方のRGBカメラが使用されるが、関連付けられていないショットとして扱われる。学習セットには20Kのユニークな画像があり、シーン分布の夕方以降に40Kのリスト(重複を含む)にシャッフルする。まず1.5Mサンプルに対して粗いモデルを学習し、次に1Mサンプルに対して細かいモデルを学習する。学習率はNYU Depthの場合と同じである。

4.3 ベースラインと比較

同じデータセットで学習させたMake3Dと、他の現在の手法の発表結果[12, 7]と、我々の手法を比較する。追加の参考として、トレーニングセット全体で計算された平均深度画像とも比較する。Make3Dは700枚の画像(シーンあたり25枚)のサブセットを用いてKITTIで学習させたが、これはシステムがこのサイズを超えるスケールアップができなかったためである。深度ターゲットは、NYUDepth開発キットのカラー化ルーチンを使用して埋めた。NYUDepthについては、795枚の画像からなる共通分布学習セットを使用した。各手法は、先行研究からのいくつかの誤差と、我々のスケール不変のメトリックを用いて評価する:

Threshold: % of y_i s.t. $\max(\frac{y_i}{y_i^*}, \frac{y_i^*}{y_i}) = \delta < thr$	RMSE (linear): $\sqrt{\frac{1}{\lvert T\rvert}\sum_{y\in T}\lvert\lvert y_i - y_i^*\rvert\rvert^2}$
Abs Relative difference: $\frac{1}{\lvert T\rvert}\sum_{y\in T}\lvert y - y^*\rvert/y^*$	RMSE (log): $\sqrt{\frac{1}{\lvert T\rvert}\sum_{y\in T}\lvert\lvert \log y_i - \log y_i^*\rvert\rvert^2}$
Squared Relative difference: $\frac{1}{\lvert T\rvert}\sum_{y\in T}\lvert\lvert y - y^*\rvert\rvert^2/y^*$	RMSE (log, scale-invariant): The error Eqn. 1

Make3Dと我々のネットワークからの予測は、入力の中央のクロップがわずかに異なることに対応することに注意してください。これらの領域の交点上で比較し、最近傍を用いて元の完全な入力解像度に予測値をアップサンプルする。アップサンプリングは、グランドトゥルースをダウンサンプリングし、出力解像度で評価するのに比べて、性能にほとんど影響を与えない[3]。

5 Results

5.1 NYU Depth

NYU Depthデータセットの結果を表1に示す。セクション4.3で説明したようにベースラインとしてデータ平均とMake3D、Karschら[7]とLadickyら[12]と比較する。(Ladickyらは、深度ラベルと意味ラベルの両方を用いて学習されるジョイントモデルを使用している)。我々のシステムは全ての指標で最高の性能を達成し、次点と比較して平均35%の相対的な利得を得た。このデータセットには、他のアプローチで使用されるデータよりも多くの例インスタンスが含まれており、関連する特徴とその関連性を学習するために効果的に活用することができる。

このデータセットは、Make3Dが行う多くの仮定、特に接地面の水平アライメントを破り、その結果、Make3Dはこのタスクの性能が比較的低い。重要なことは、我々の手法は、スケール依存とスケール不変の両方のメトリクスでそれを改善し、我々のシステムがより良い平均だけでなく、より良い関係も予測できることを示していることである。

定性的な結果は、図4の左側に示されており、スケール不変MSEによって上から下へソートされている。誤差測定では、微細なネットワークは改善されないが、その効果は深度マップではっきりと確認できる – 表面境界はよりシャープな遷移を持ち、局所的な詳細に整列する。しかし、いくつかのテクスチャエッジも含まれることがある。図3は、Make3Dと、λ=0とλ=0.5を用いた損失で学習したネットワークの出力を比較したものである。λ = 0.5では、λ = 0よりも数値的な利得は観察されなかったが、より詳細な出力ではわずかな質的な改善が見られた。

NYUDepthでは、アップサンプリングとダウンサンプリングの対数RMSEはそれぞれ0.285対0.286、スケール不変RMSEは0.219対0.221である。NYUDepthではネットワーク領域の86%、Make3Dの100%、KITTIではネットワークの100%、Make3Dの82%が交差している。

	Mean	Make3D	Ladicky*&al*	Karsch*&al*	Coarse	Coarse + Fine	
threshold $\delta < 1.25$	0.418	0.447	0.542	–	**0.618**	0.611	higher
threshold $\delta < 1.25^2$	0.711	0.745	0.829	–	**0.891**	0.887	is
threshold $\delta < 1.25^3$	0.874	0.897	0.940	–	0.969	**0.971**	better
abs relative difference	0.408	0.349	–	0.350	0.228	**0.215**	
sqr relative difference	0.581	0.492	–	–	0.223	**0.212**	lower
RMSE (linear)	1.244	1.214	–	1.2	**0.871**	0.907	is
RMSE (log)	0.430	0.409	–	–	**0.283**	0.285	better
RMSE (log, scale inv.)	0.304	0.325	–	–	0.221	**0.219**	

表1:NYUDepthデータセットでの比較

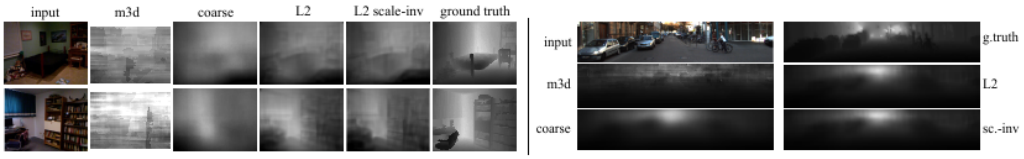


図3: Make3D、l_2損失で学習した我々の手法($\lambda = 0$)、l_2損失とスケール不変損失の両方で学習した我々の手法($\lambda = 0.5$)の定性的比較。

5.2 KITTI

次に、KITTI走行データセットでの結果を検証する。ここで、Make3Dベースラインは、水平方向に整列した画像で構成され、データセットに適しており、比較的良好な結果を達成する。それでも、我々の方法は、すべてのメトリクスで、平均31%の相対的な利得で、それを上回る。同様に重要なことは、スケール依存RMSE誤差とスケール不変RMSE誤差の両方が25%増加し、予測された構造が大幅に改善されることを示していることである。ここでも、微細なネットワークは、誤差メトリクスでは粗いネットワークよりもあまり改善されないが、定性的な出力では両者の違いが見られる。

図4の右側は、予測値の例を、やはり誤差でソートしたものである。ファインスケールネットワークは、特に道路端付近で、よりシャープな遷移を生成する。しかし、その変化はやや限定的である。これは、回転スキャナのセットアップに起因する、深度マップとトレーニングデータの入力の間の補正されていないアライメントの問題が原因であると思われる。これにより、エッジは真の位置から切り離され、ネットワークはよりランダムな配置で平均化される。図3は、予想通り、Make3Dがこのデータではるかに良い性能を発揮していることを示しているが、この場合、スケール不変誤差を損失として使用することは、ほとんど効果がないようである。

	Mean	Make3D	Coarse	Coarse + Fine	
threshold $\delta < 1.25$	0.556	0.601	0.679	**0.692**	higher
threshold $\delta < 1.25^2$	0.752	0.820	0.897	**0.899**	is
threshold $\delta < 1.25^3$	0.870	0.926	**0.967**	**0.967**	better
abs relative difference	0.412	0.280	0.194	**0.190**	
sqr relative difference	5.712	3.012	1.531	**1.515**	lower
RMSE (linear)	9.635	8.734	7.216	**7.156**	is
RMSE (log)	0.444	0.361	0.273	**0.270**	better
RMSE (log, scale inv.)	0.359	0.327	0.248	**0.246**	

表2:KITTIデータセットでの比較。

6 Discussion

単一の画像から奥行き推定値を予測することは困難な作業である。しかし、グローバルビューとローカルビューの両方からの情報を組み合わせることで、それなりにうまく実行することができる。我々のシステムは、2つのディープネットワークを使用することでこれを達成する。1つはグローバルな深度構造を推定するもので、もう1つは局所的により細かい解像度で深度構造を洗練させるものである。我々は、NYU DepthとKITTIデータセットにおいて、生データの完全な分布を効果的に活用し、このタスクで新たな最先端を達成した。

今後の研究では、表面法線のような3D形状情報をさらに取り込むために、我々の方法を拡張する予定である。法線マップ予測における有望な結果はFouheyらによってなされており[2]、深度マップとともにそれらを統合することで、全体的な性能が向上する[16]。また、より細かいスケールのローカルネットワークを繰り返し適用することで、深度マップを元の完全な入力解像度に拡張することを期待している。

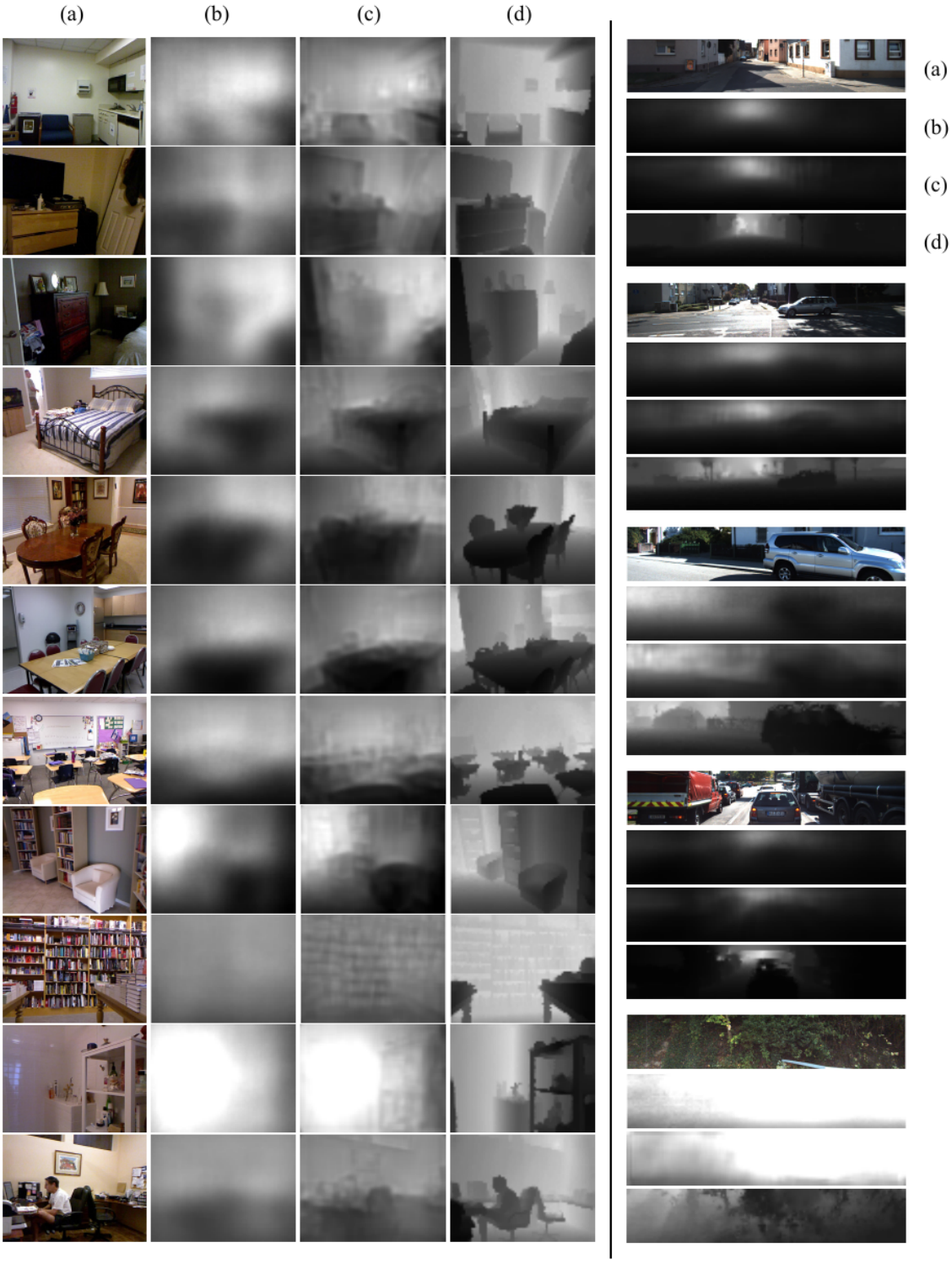


図4:我々のアルゴリズムによる予測例。左がNYUDepth、右がKITTI。各画像について、(a)入力、(b)粗いネットワークの出力、(c)細かいネットワークの洗練された出力、(d)グランドトゥルースを示す。ファインスケールネットワークは、粗いスケールの入力を編集し、オブジェクトの境界や壁のエッジなどの詳細との整合性を高める。例題はベスト(上)からワースト(下)にソートされている。

References

[1] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[2] D. F. Fouhey, A. Gupta, and M. Hebert. Data-driven 3d primitives for single image understanding. In *ICCV*, 2013.

[3] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[4] R. Hadsell, P. Sermanet, J. Ben, A. Erkan, M. Scoffier, K. Kavukcuoglu, U. Muller, and Y. LeCun. Learning long-range vision for autonomous off-road driving. *Journal of Field Robotics*, 26(2):120–144, 2009.

[5] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[6] D. Hoiem, A. A. Efros, and M. Hebert. Automatic photo pop-up. In *ACM SIGGRAPH*, pages 577–584, 2005.

[7] K. Karsch, C. Liu, S. B. Kang, and N. England. Depth extraction from video using non-parametric sampling. In *TPAMI*, 2014.

[8] K. Konda and R. Memisevic. Unsupervised learning of depth and motion. In *arXiv:1312.3429v2*, 2013.

[9] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[10] A. Levin, R. Fergus, F. Durand, and W. T. Freeman. Image and depth from a conventional camera with a coded aperture. In *SIGGRAPH*, 2007.

[11] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. Freeman. Sift flow: dense correspondence across difference scenes. 2008.

[12] M. P. Lubor Ladicky, Jianbo Shi. Pulling things out of perspective. In *CVPR*, 2014.

[13] R. Memisevic and C. Conrad. Stereopsis via deep learning. In *NIPS Workshop on Deep Learning*, 2011.

[14] J. Michels, A. Saxena, and A. Y. Ng. High speed obstacle avoidance using monocular vision and reinforcement learning. In *ICML*, pages 593–600, 2005.

[15] A. Saxena, S. H. Chung, and A. Y. Ng. Learning depth from single monocular images. In *NIPS*, 2005.

[16] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3-d scene structure from a single still image. *TPAMI*, 2008.

[17] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47:7–42, 2002.

[18] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[19] F. H. Sinz, J. Q. Candela, G. H. Bakır, C. E. Rasmussen, and M. O. Franz. Learning depth from stereo. In *Pattern Recognition*, pages 245–252. Springer, 2004.

[20] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: Exploring photo collections in 3d. 2006.

[21] K. Yamaguchi, T. Hazan, D. Mcallester, and R. Urtasun. Continuous markov random fields for robust stereo estimation. In *arXiv:1204.1393v1*, 2012.